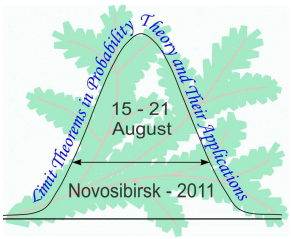

Parallel server queueing systems in the Halfin-Whitt heavy traffic regime

David Gamarnik

A parallel server queueing system model is used in a variety of applications including computer networks, call centers and health care management. Understanding the behavior of this system in the heavy traffic setting when the number of servers is large is a very challenging problem. While a lot is known in the special case of exponentially distributed holding times, starting with the classical work of Halfin and Whitt in 1981, far less is known in the non-exponential case. This is unfortunate since the real life data, for example the number of days spent by patients in a hospital, often suggests distributions far from exponential.

We will present a recent progress in understanding the steady state behavior of a parallel server queueing system in the heavy traffic regime when the holding time distribution is arbitrary. Specifically, we obtain a surprisingly simple upper bound on the limiting tail distribution of the queue length in terms of the stochastic primitives of the model. In special cases we establish the tightness of this bound.

Joint work with David Goldberg (MIT).